DOI: 10.61782/fa.2023.0412

forum acusticum 2023

ADAPTATION TO ALTERED INTERAURAL TIME DIFFERENCES IN A VIRTUAL REALITY ENVIRONMENT

Pierre Guiraud[1*] **Kevin Sum**[1] **Niels Henrik Pontoppidan**[2] **Katarina C. Poole**[1] **Lorenzo Picinali**[1]

[1] Dyson School of Design Engineering, Imperial College London, London, UK.
[2] Eriksholm Research Centre, Denmark

ABSTRACT

Interaural time differences (ITDs) are important cues for determining the azimuth location of a sound source and need to be accurately reproduced, in a virtual reality (VR) environment, to achieve a realistic sense of sound location for the listener. ITDs are usually included in head related transfer functions (HRTFs) used for audio rendering, and can be individualised to match the user's head size (e.g. longer ITDs are needed for larger head sizes). In recent years, studies have shown that it is possible to train subjects to adapt and improve their performance in sound localisation skills to non-individualized HRTFs. The analysis of such improvements has focused mainly on adaptation to monoaural spectral cues rather than binaural cues such as ITDs. In this work listeners are placed in a VR environment and are asked to localise the source of a noise burst in the horizontal plane. Using a generic non-individualized HRTF with its ITD modified to match the head size of each participant, test and training phases are alternated, with the latter providing continuous auditory feedback. The experiment is then repeated with ITDs simulating larger (150%) and smaller (50%) head sizes. Comparing localisation accuracy before and after training, it is observed that while training seems to improve sound localisation performance, this varies according to the simulated head size and target location.

Keywords: *ITD, training, VR, HRTF*

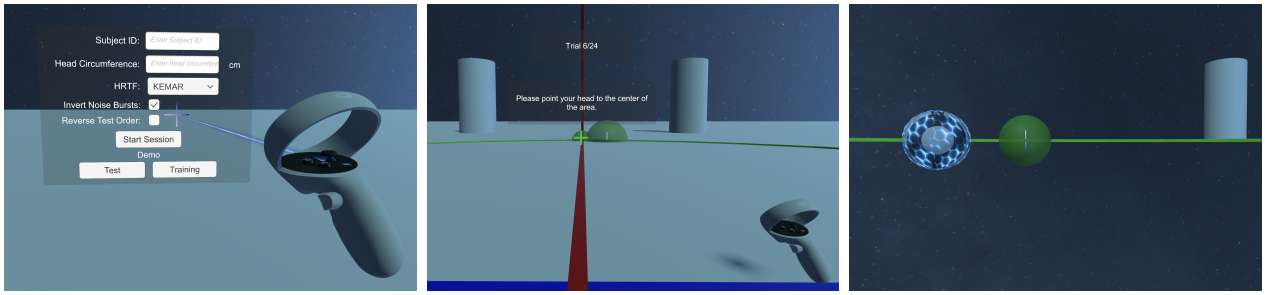
**Corresponding author*: p.guiraud@imperial.ac.uk.

1. INTRODUCTION

In order to locate the source of sounds in space, the human auditory system uses monaural spectral cues, like the direction dependent filtering of the pinnae, and binaural cues, such as the interaural time and level differences (ITDs and ILDs) [1]. Spatial hearing is essential for segregating listening targets from background noise and/or multiple speakers, also known as the *cocktail party problem* [2]. Localisation of individual sound sources and thereby perceptual segregation of these sources, whether target and/or masker, can aid in increasing speech intelligibility. However, these scenarios are particularly challenging for hearing impaired people [3] where low-frequency hearing loss limits access to ITD cues and high-frequency hearing loss limits access to monaural spectral cues [4].

With the appearance of virtual reality (VR) there has been an increase in research interest in generating artificial spatial cues. A popular method for spatial audio generation uses head related transfer functions (HRTFs) [5]. HRTFs describe the angular dependent acoustic modifications induced by a person ears and head on sound in a free-field environment. HRTFs are unique to each individual due to the unique shape and size of a listeners pinnae, head and torso. Hence the use of generic HRTFs can lower sound localisation performance for speech in noise situations [6]. However HRTFs are challenging to acoustically measure as they require specialised equipment and lengthy measurements [7]. Therefore, in order for accurate spatial audio reproduction to be more commonplace, an alternative to individual measurements must be found.

In recent years, it has been shown that whilst performance to non-individual HRTFs is initially poor in terms of sound localisation, listeners can adapt to the non-



(a) (b) (c)

Figure 1: Visuals of the virtual test environment. (a) displays the starting screen where the participant's information and the test conditions are entered. (b) shows a test example. The green ball in the horizontal plane is the pointer of the controller and the cross is the head orientation of the participant. The cross is here green as it is facing forward at the intersection of the horizontal (green) and vertical (red) line. (c) shows an example of a source to localise when it is visible. The controller's pointer is shown next to it.

individual spatial cues given enough exposure and training [8–10]. However, training on non-individual HRTFs has been focused on the adaptation to monoaural spectral cues rather than binaural ones, such as ITDs, which are essential to accurately determine the azimuth location of a sound source [11].

In this work, the sound localisation accuracy in the horizontal plane is investigated when using a generic HRTF with modified ITD cues. Participants are set in a VR environment and are asked to localise a sound source in the horizontal plane. After an initial baseline localisation test, they are trained for several minutes on the localisation task using auditory and visual feedback before performing the test again. This is then repeated with a modified head size (50 or 150 %). The influence of various effects, such as training, head size, training order or target location, are analysed using t-tests [12] and generalised linear mixed effects models (GLME) [13].

2. EXPERIMENTAL SETUP

2.1 Tools

The experiment is conducted on an Oculus Quest 2 with Sennheiser HD 650 headphones without headphone equalisation. The application is developed on Unity and the audio spatialisation is performed using the 3D Tune-in toolkit (3DTI) [14]. The non-individual HRTF used in this study is from a KEMAR mannequin mounting *large* from the SONICOM HRTF dataset [15]. The ITD was removed from the HRTF and then added back with the 3DTI spatialiser plugin, using the participant's head circumference as the input.

2.2 Procedure

Participants entering the experiment were asked the following series of question: their age range, how often they play video games and how familiar they are with immersive audio. Their head size is measured and used for sound spatialisation in the experiment. When starting the experiment, the screen seen in Fig. 1a appears. The participant enters their name, head size and selects the KEMAR HRTF. A pre-defined tick box allows to inverse the order of modified ITD during the test, see below. The participant can then start the experiment.

The experiments can be summarised in three blocks as described in Table 1. The first block consists of a localisation task, 6 min of training, and another localisation task using the participant's true head size. This block serves as reference and is useful for participants to get familiar

Table 1: VR experiment block summary for various head sizes. "L" stands for localisation task with 24 targets and "T" for a 6 min training task. The experiment always starts with 100% head size.

	Tasks order
Head Size: 100%	L-T-L
Head Size: 150%	L-T-T-L
Head Size: 50%	L-T-T-L

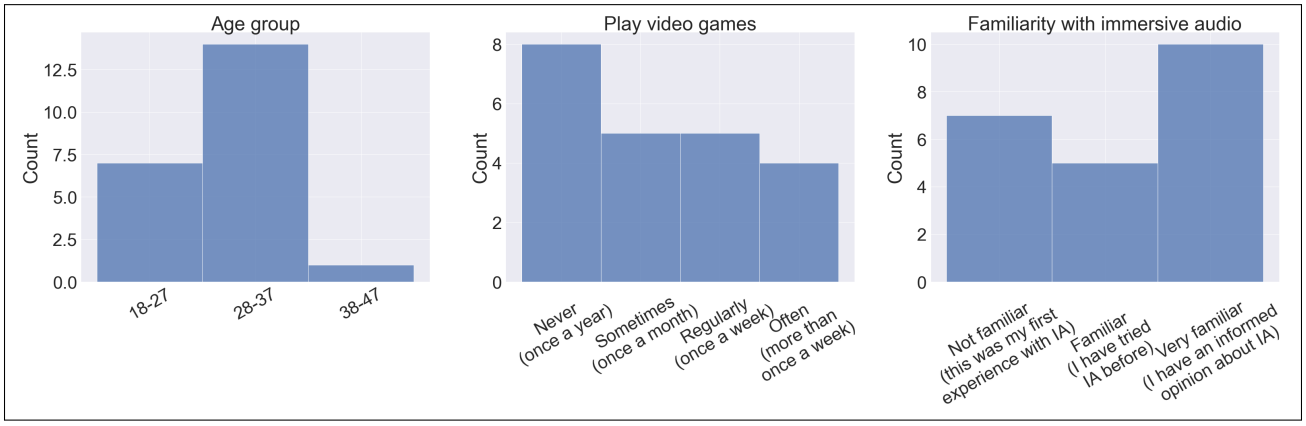


Figure 2: Participants screening survey results. Display age groups, video game practice and familiarity with immersive audio (IA).

with the test and the VR space. In the second and third blocks, participant's head size are modified to either 50 or 150 percent of their original head size. In each of these block they perform first a localisation task, then do 12 min of training, and finish with another localisation task as described in Table 1. The order in which those blocks are presented was randomised between participants using the aforementioned tick box. The total length of the test is between 40 min and 1 h.

2.3 Localisation task

During the localisation task, participants are asked to estimate the location of an invisible sound source. They are first asked to align their head with the virtual spherical rig. As seen in Fig. 1b, a cross showing them where their head is pointing toward is provided to help them align. Once aligned, a 250 ms long noise burst is played at 0 degrees elevation and at a random azimuth angle in the 180 degrees frontal hemifield, with a float degree resolution. The controller's pointer is locked on the horizontal plane as observed in Fig. 1b. Participants must then point toward the localisation of the source and use the trigger to confirm the predicted target location. If the participant needs to listen to the short burst again, it will be repeated once after 3 s provided that their head is still positioned forwards. This procedure is repeated for a total of 24 locations. Verification that the head is aligned with the spherical rig is performed with every new target. No performance feedback is given during this task.

2.4 Training task

During the training task, participants are asked to determine the location of invisible targets. To do this, the controller's pointer provides auditory feedback in relation to the target's location. The short burst used in the localisation task is now repeated continuously. The repetition rate and duration of the burst are altered as the controller's pointer moves closer or further away from the invisible target. The lowest repetition rate and longer duration correspond to the target location. Similarly to the localisation task, participants are required to keep their head facing forward while searching otherwise, if the head faces away, the bursts become rapidly inaudible. Once the participants select the location of the predicted sound and this is less than 20 degrees away from the target location, it is made visible and participants are asked to point their head towards the target location before being allowed to proceed to the next trial. Figure 1c shows the tracker/pointer next to the visible identified target. Similarly as in the localisation task, a cross showing where the participant's head is pointing is provided as visual cue. This is repeated for a total of 6 min.

To keep participants engaged, the training task is gamified using a scoring system. For each trial, participants gain 10 points for each target found, 10 more points if they find the target on their first try, and 10 extra points for accuracy, answer less than 10 degrees away from target.

Table 2: Model summary for generalised mixed effects linear regression on angular error for fixed effects: different head sizes, before and after training and whether or not the target is within field of view of $\pm 45°$ (fov). Data obtained on 22 participants. Categorical reference used: head size $100\,\%$, before training, experiment order 100-50-150, outside fov.

Fixed effects	Estimate	SE	tStat	pValue
Intercept	17.64	2.25	9.97	< 0.001*
Head size: 150%	10.82	1.62	6.66	< 0.001*
Head size: 50%	11.35	1.62	6.99	< 0.001*
After training	0.26	1.62	0.16	0.87
Reverse experiment order	0.16	2.78	0.06	0.95
Within fov	0.32	1.62	0.20	0.84
Head size: 150% $\times$ After training	-6.48	2.29	-2.83	0.005*
Head size: 50% $\times$ After training	-5.60	2.28	-2.46	0.01*
Head size: 150% $\times$ Within fov	-10.31	2.29	-4.50	< 0.001*
Head size: 50% $\times$ Within fov	-15.61	2.29	-6.81	< 0.001*
After training $\times$ Within fov	-5.62	2.29	-2.46	0.01*
Head size: 150% $\times$ After training $\times$ Within fov	9.36	3.24	2.89	0.004*
Head size: 50% $\times$ After training $\times$ Within fov	11.16	3.23	3.45	< 0.001*

ID random effect intercept standard deviation estimate: 6.33
R^2 adjusted: 0.15
Degrees of freedom: 3310
*: $p < 0.05$

3. RESULTS AND DISCUSSION

3.1 Participants

22 participants took part in this experiment. 7 of them are women, 4 were left handed and the head circumference at ear level ranged from $57\,\mathrm{cm}$ to $63\,\mathrm{cm}$ with a mean size of $60.5\,\mathrm{cm}$. Upon entering the experiment some questions were asked as described in Section 2.2. The distribution of the results are seen in Fig. 2. 11 people did the test with the head size order $100\,\%$ to $50\,\%$ to $150\,\%$, and 11 people in the order $100\,\%$ to $150\,\%$ to $50\,\%$, as observed in Table 3.

3.2 Generalised mixed-effect model analysis

To investigate the effect of training on the aggregated results, a generalized linear mixed-effects model (glme) fitted by maximum probable loss is used to predict angular error. With Wilkinson notation, the formula used is

$$angularError \sim headSize * training * fov + headSizeOrder + (1|ID), \quad (1)$$

with fixed effects being the different head sizes (*headSize)*, results before and after training (*training)*, whether the target is within or outside field of view (*fov*) and the order in which the head size were presented during the test as seen in Table 3 (*headSizeOrder*). Participants ID are used as random effect (*ID)*. Fov is defined as the target being within $\pm 45°$. This allows for an even split of the data while being coherent with the true VR field of view. Notably, a $\pm 53°$ fov split corresponding to the limit of an unfamiliar ITD with $150\,\%$ head size using the Woodworth formula [16] was also investigating with no significant difference compared to $\pm 45°$. Results of the glme are seen in Table 2 using as categorical reference: "*head size* $100\,\%$", "*before training*", "*outside fov*" and "*experiment order 100-50-150*".

A fixed effect is considered to have an impact if its pValue is below $5\,\%$. It is seen that training alone does not lead to statistical significant results as its pValue seen in Table 2 is above this $5\,\%$ threshold. The order of head sizes in which the experiment was perform is also not significant. The glme showed significant main effects of change of head size and target location, as well as significant interactions between head size, target location and training. To pull the effect of training in those specific cases, additional glme's for each *headSize* and *fov* are run. The resulting six scenarios are observed in Table 4. Train-

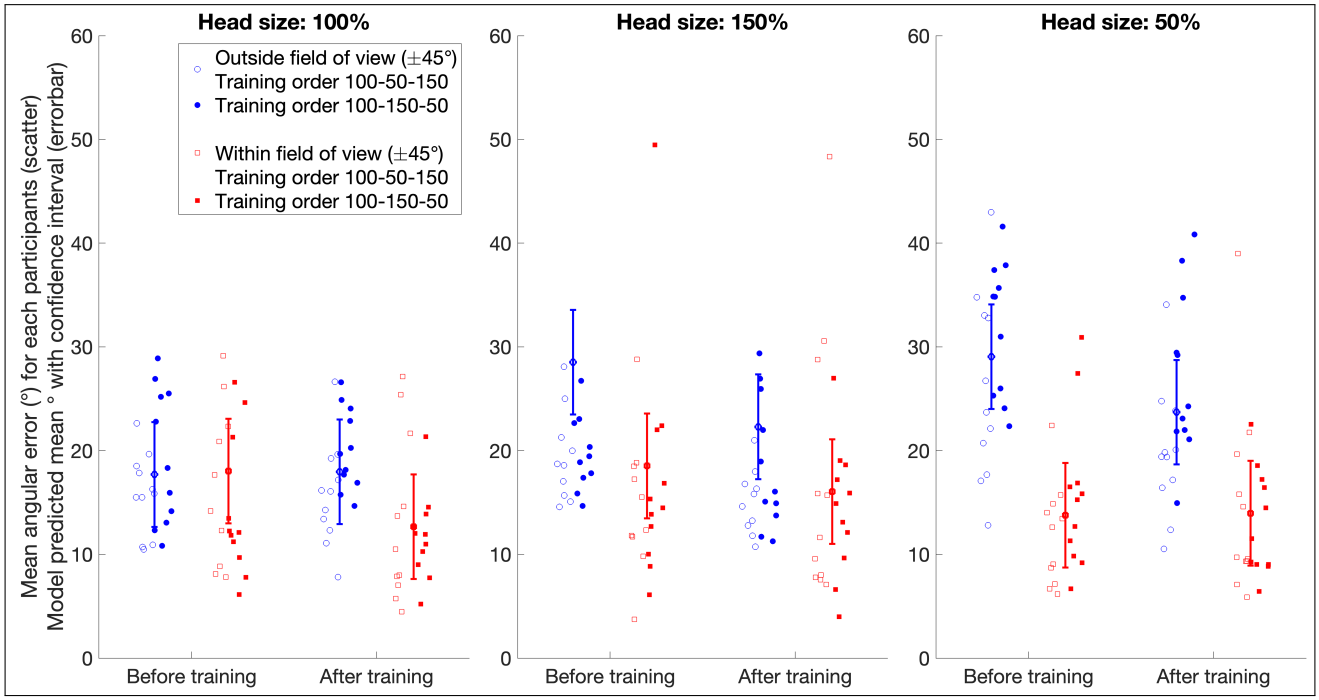


Figure 3: From left to right, graphs show mean angular error before and after training with head size of 100%, 150% and 50%. Blue circles show mean individual experimental results per participant for localisation target outside the field of vision (FOV) $\pm 45°$. Red squares represent similar results for localisation target within FOV. Empty and filled points relate to the test order of the participant. Corresponding error bars show the predicted mean across participants and the 95% confidence interval using the complete glme model.

ing with a normal head size suggest a significant reduction of the angular error by $5.66°$ when the target is within fov. This is not the case for target outide fov. This effect can be attributed to participants getting used to the virtual environment and the headset as this is the first test they perform. With a larger head size, training improves localisation for targets inside and outside fov by $2.60°$ and $6.02°$ respectively. The more pronounced effect for targets outside fov can be attributed by the broadened sense of sound directionality due to larger ITD cues from increased lateralisation of the target sound source. Targets outside fov are localised using mainly unfamiliar ITDs (limit of known ITD at $\pm 53°$) making for a challenging task. Training provides participants with some reference and an awareness of the new soundscape improving more drastically the results than for targets within fov. Similar observations are made with smaller head size for targets outside the fov with improvements of $5.35°$. For targets within fov, no significant improvement is observed. With smaller ITDs, the sense of sound localisation is reduced and most sounds feel like coming from the front. When a target is with the fov the participant, a reduced ITD does not hinder their ability to localise and training does not help. Additionally in Table 4, the influence of the participants' familiarity with video games and immersive audio are added as fixed effects. While those effects are significant in some cases, the irregularity of the results and the reduced number of participants in those subgroups prevents us from drawing significant conclusions. The effect of age could not be investigated due imbalanced data as observed in Fig. 2.

To visualise the effect of head size and target location, predictions of angular error for each participant and for each condition are made using the full glme model define in Eq. 1. Results are presented in Fig. 3. The error bars correspond to the predicted means and their $95\,\%$ confidence interval using the glme model. Scatter points correspond to the experimental mean result of each participant.

Blue circles correspond to mean results for targets outside fov, and red squares for target within fov. Empty and filled points distinguish participants depending on head size test order

Corroborating Table 2, training shows an improvement for normal head size within fov, smaller head size outside fov, and larger head size in both cases. Considering the results of normal head size after training as the best achievable performance, it is observed that using a modified head size lead to poorer performance even after training. Notably, only the results of a small head size with target within fov achieve comparable performance. Smaller ITDs then do not change the localisation accuracy when the target is up to a certain angle but will still "shrink" the soundscape for sources located on the sides.

Table 3: pValue of related t-test performed on each participant localisation task before and after training with different head size.

ID	Head size: 100 %	Head size: 150 %	Head size: 50 %
0^a	0.91	0.46	0.24
1^a	0.31	0.22	0.13
2^a	0.07	0.79	0.08
3^b	0.82	0.07	0.63
4^a	0.75	0.33	0.21
5^a	0.44	0.35	0.22
6^b	0.16	0.20	0.69
7^a	0.68	$< 0.001^{*\dagger}$	0.78
8^b	0.38	$0.79^{\dagger}$	0.57
9^a	0.94	0.40	0.12
10^a	0.61	0.11	0.93
11^b	$< 0.001^*$	0.83	0.03^*
12^b	0.15	0.15	0.83
13^a	0.17	0.56	0.77
14^a	0.59	0.62	0.84
15^b	0.61	0.01^*	$< 0.001^*$
16^b	0.08	0.84	0.59
17^a	0.49	0.42	0.40
18^b	$0.58^{\dagger}$	0.60	0.50
19^b	0.99	0.04^*	0.08
20^b	0.03^*	0.21	0.83
21^b	0.50	0.12	0.89

*: $p < 0.05$
†: outlier result
a: test order 100 % - 150 % - 50 %
b: test order 100 % - 50 % - 150 %

3.3 Effect of individual training

In order to assess the individual effect of training with different head sizes, two-tails related t-test are performed on localisation task performances for each participants. The angular error difference between the target true and estimated position is investigated before and after training and results are displayed in Table 3.

The pValue of the t-test is below the 5 % threshold only 7 times out of the total of 66 cases. It is then not possible to conclude that training alone can help horizontal sound localisation when head size, and so ITD, is largely modified.

Notably, results are considered to be outliers if the mean angular error before or after training is above 70°. Those discrepancies are attributed to experimental error. In particular, the result of participant ID 7 with head size 150 % is observed to be an outlier. Its statistical significant training is then not relevant.

Taking this into account, only 9 % of the cases showed significant difference after training. While this gives an estimation of individual effect, more trials are needed to increase statistical power and the relevance of this result. The glme performed on the whole population in Section 3.2 being a more robust analysis. The individual data can be downloaded at the following GitHub address [17].

4. CONCLUSIONS

In this work listeners in a VR setting are being trained to locate an invisible sound source in the horizontal plane using a generic HRTF with various head sizes. More specifically, the ITD component of the HRTF is altered simulating head sizes of 100 %, 150 % and 50 % of participant's original measure. The influence of the head size and the target location (within or outside field of view) before and after training is using t-tests and/or generalised linear mixed effect models (glme).

Individual t-tests for each participant do not show a statistically significant effect of training, regardless of head size. However, a glme model on the aggregated results shows that with a modified ITD, training does help increase localisation accuracy especially for targets outside the field of view (fov). Nonetheless, the most accurate results in the sound localisation task are still found for a normal head size, and only targets within fov using the smaller head size reached similar performances after training. Notably, the order in which the test presented the

head sizes did not significantly affect performance, and it is still unclear how the familiarity with video games and immersive audio influences performance in sound localisation.

It has been shown that training improves sound localisation performances in the horizontal plane when using a generic HRTF with modified ITDs. While performance with unrealistic ITDs after training does not reach those of a normal head size, this may be due to the short duration of the training task. Future work may focus on exploring an increase of the training time, as well as for how long the localisation skills with altered ITDs are retained. Furthermore, simulations involving also ILD alterations for increased and/or decreased head sizes might be implemented. Lastly, in addition to sound localisation, this experimental protocol could also be used to assess for other spatial perception metrics in VR, like externalisation and immersiveness.

5. REFERENCES

[1] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[2] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," 2000.

[3] T. V. den Bogaert, T. J. Klasen, M. Moonen, L. V. Deun, and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *The Journal of the Acoustical Society of America*, vol. 119, 2006.

[4] T. Neher, T. Behrens, S. Carlile, C. Jin, L. Kragelund, A. S. Petersen, and A. V. Schaik, "Benefit from spatial separation of multiple talkers in bilateral hearing-aid users: Effects of hearing loss, age, and cognition," *International Journal of Audiology*, vol. 48, 2009.

[5] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," 2020.

[6] M. Cuevas-Rodriguez, D. Gonzalez-Toledo, A. Reyes-Lecuona, and L. Picinali, "Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment," *The Journal of the Acoustical Society of America*, vol. 149, 2021.

[7] L. Picinali and B. F. Katz, "System-to-user and user-to-system adaptations in binaural audio," in *Sonic Interactions in Virtual Environments*, pp. 115–143, Springer International Publishing Cham, 2022.

[8] S. Carlile, K. Balachandar, and H. Kelly, "Accommodating to new ears: The effects of sensory and sensory-motor feedback," *The Journal of the Acoustical Society of America*, vol. 135, 2014.

[9] M. A. Steadman, C. Kim, J. H. Lestang, D. F. Goodman, and L. Picinali, "Short-term effects of sound localization training in virtual reality," *Scientific Reports*, vol. 9, 12 2019.

[10] P. Stitt, L. Picinali, and B. F. Katz, "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," *Scientific reports*, vol. 9, no. 1, p. 1063, 2019.

[11] L. L. Young Jr and J. Levine, "Time-intensity trades revisited," *The Journal of the Acoustical Society of America*, vol. 61, no. 2, pp. 607–609, 1977.

[12] M. Xu, D. Fralick, J. Z. Zheng, B. Wang, X. M. Tu, and C. Feng, "The differences and similarities between two-sample t-test and paired t-test," *Shanghai Archives of Psychiatry*, vol. 29, 2017.

[13] S. T. Gries, "(generalized linear) mixed-effects modeling: A learner corpus example," *Language Learning*, vol. 71, 2021.

[14] M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, and A. Reyes-Lecuona, "3d tune-in toolkit: An open-source library for real-time binaural spatialisation," *PLoS ONE*, vol. 14, 3 2019.

[15] I. Engel, R. Daugintis, T. Vicente, A. O. T. Hogg, J. Pauwels, A. J. Tournier, and L. Picinali, "The sonicom hrtf dataset," *Journal of the Audio Engineering Society - accepted for publication in December 2022*, 2022.

[16] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, 1998.

[17] `https://github.com/Audio-Experience-Design/ForumAcusticum-ITDinVR-data-Guiraud2023.git`. Accessed: 2023-04-24.

Table 4: Generalised mixed effects model for cases with different head sizes and whether or not the target is within field of view (fov). Data obtained on 22 participants. Categorical reference used: before training, not familiar with immersive audio (IA). The familiarity with video games was regrouped in a binary outcome to increase the number of data of those subgroups.

Fixed effects	Estimate	SE	tStat	pValue
Head 100%, outside fov				
Intercept	19.49	1.47	13.23	< 0.001*
After training	0.42	1.38	0.30	0.76
Plays video games	-0.02	1.45	-0.01	0.99
Familiar with IA	2.98	1.94	1.54	0.12
Very familiar with IA	-5.53	1.61	-3.43	< 0.001*
Head 100%, within fov				
Intercept	22.94	3.03	7.57	< 0.001*
After training	-5.66	1.34	-4.23	< 0.001*
Plays video games	-7.90	3.31	-2.38	0.02*
Familiar with IA	-4.50	4.47	-1.01	0.31
Very familiar with IA	-1.47	3.67	-0.40	0.69
Head 150%, outside fov				
Intercept	16.94	8.00	2.12	0.03*
After training	-6.02	1.65	-3.64	< 0.001*
Plays video games	11.37	8.89	1.28	0.20
Familiar with IA	30.25	12.01	2.52	0.01*
Very familiar with IA	-0.27	9.88	-0.03	0.98
Head 150%, within fov				
Intercept	20.20	3.62	5.62	< 0.001*
After training	-2.60	1.26	-2.07	0.04*
Plays video games	0.74	3.96	0.19	0.85
Familiar with IA	6.82	5.34	1.28	0.20
Very familiar with IA	-7.74	4.39	-1.76	0.07
Head 50%, outside fov				
Intercept	26.71	2.76	9.69	< 0.001*
After training	-5.35	1.51	-3.53	< 0.001*
Plays video games	3.00	2.97	1.01	0.31
Familiar with IA	1.80	4.02	0.45	0.65
Very familiar with IA	1.41	3.30	0.43	0.67
Head 50%, within fov				
Intercept	13.85	1.73	11.44	< 0.001*
After training	0.05	0.98	0.05	0.96
Plays video games	-3.49	1.85	-1.88	0.06
Familiar with IA	-6.69	2.50	-2.68	0.007*
Very familiar with IA	-6.64	2.06	-3.22	0.001*

ID random effect intercept standard deviation estimate: 0.002, 6.73, 19.56, 8.37, 5.63, 5.13.
R^2 adjusted: 0.039, 0.21, 0.62, 0.32, 0.089, 0.14.
Degrees of freedom: 549, 550, 547, 549, 555, 543.
*: $p < 0.05$